Combining productivity and predictability for measuring morphological complexity

Ximena Gutierrez-Vasques
xim@unam.mx
Víctor Mijangos
victor.mijangosc@gmail.com

1. Which level of language is addressed

In this work we address the morphology level. Languages of the world have different word production processes. Therefore, the amount of semantic and grammatical information encoded at the word level, may vary significantly from language to language. In this sense, it is important to quantify the morphological richness of languages and how it varies depending on their linguistic typology (Baerman, Brown and Corbett, 2010).

2. What exactly is measured

Conceptualizing and quantifying linguistic complexity is not an easy task, many quantitative and qualitative dimensions must be taken into account (Miestamo, 2008). In general terms, the complexity of a system could be related to the number and variety of elements, but also to the elaborateness of their interrelational structure (Simon, 1996: 183; Sinnemäki, 2011: 16).

In the case of morphological complexity, several corpus-based methods are successful in capturing the number and variety of the morphological elements of a language by measuring the distribution of words over a corpus. However, they may not capture other complexity dimensions like the predictability of the internal structure of words. There can be cases where a language is considered complex because it has a rich morphological productivity, i.e., great number of morphs can be encoded into a single word. However, the combinatorial structure of these morphs in the word formation process can have less uncertainty than other languages, i.e., more predictable.

We would like to quantify the morphological complexity by measuring the type and token distributions over a corpus, but also by taking into account the predictability of the subword sequences within a word (Montermini and Bonami, 2013).

We conjecture that the predictability of the internal structure of words reflects the difficulty of producing novel words given a set of lexical items (stems, suffixes or morphs). We take as an inspiration the statistical language models used in NLP, which are a useful tool for calculating the probability of any sequence of words in a language. However, we adapt this notion to the subword level. Information theory based measures (entropy) can be used to estimate the predictiveness of these models.

To sum up, we approach the morphological complexity by combining two different measures over parallel corpora: a) the type/token relationship (TTR); and b) the entropy of a subword language model as a measure of predictability

3. How the measure is calculated

Type/token relationship (TTR): This is a very straight forward measure that can be taken as an indicator of morphological complexity (Kettunen, 2014). The idea is that a very productive morphological system will produce a wide variety of word forms, this will be reflected in a higher TTR. The TTR over a corpus can be calculated as:

$$TTR = \frac{\#types}{\#tokens}$$

Where *#types* are the different word types in the corpus (vocabulary size), and *#tokens* is the total number of word tokens in the corpus.

Entropy rate of a subword language model: We propose this measure as an indicator of predictability of the internal structure of words in a language. We conjecture that morphological processes that are irregular/suppletive, unproductive, etc. will increase the entropy of a model that predicts the probability of sequences of morphs/subword units within a word.

In order to do this, we estimate a stochastic matrix P, where each cell contains the transition probability between two sub-word units in that language (see Table 1). These probabilities are estimated using the corpus and a neural language model that we will describe below.

	#ca	*cat*	*ats*	*ts$*
#ca	0.01	0.06	0.07	0.33
cat	0.9	0.04	0.05	0.22
ats	0.06	0.78	0.05	0.23
ts$	0.03	0.12	0.83	0.22

Table 1. Example of a transition matrix for the word *cat* using trigrams of characters

Regarding to the sub-word units, it would be difficult to perform morphological segmentation to all the languages in the dataset in order to extract "morphs". Instead of this, we focused on fixed-length sequences of characters (*n*-grams). There is evidence that trigrams of characters capture information about partial order and they encode morphological characteristics of the word (Baayen, Chuan and Blevins, 2018; Vania, 2018). In fact, Baayen, Chuan and Blevins (2018) use the term "triphone" for this unit. Since, in this dataset, there are languages with syllabic writing systems, we also took into consideration unigrams (characters).

We calculate a stochastic matrix P as follows:

$$P = p_{ij} = p(w_j|w_i) \quad (1)$$

Where w_i and w_j are *n*-grams. We used a neural probabilistic language model to estimate a probability function. Our model was obtained using a feedforward neural network; this network gets trained with pairs of consecutive *n*-grams that appear in the same word. Once the network is trained we can retrieve from the output layer the probability p_{ij} for any pair of ngrams.

This architecture is based on Bengio (2003); however, we used character *n*-grams instead of words. The network comprises the following layers: 1) an input layer of one-hot vectors representing the *n*-grams; 2) an embedding layer; 3) a hyperbolic tangent hidden layer; 4) finally, an output layer that contains the conditional probabilities obtained by a Softmax function (Figure 1).

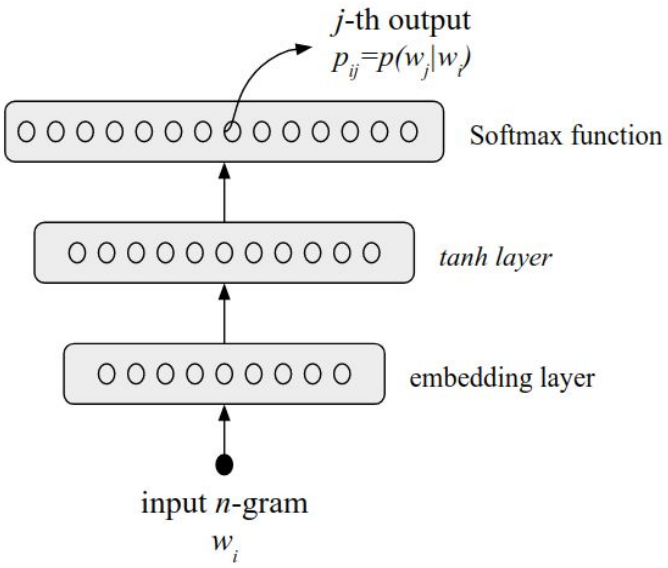

Figure 1. Neural probabilistic language model architecture (Bengio, 2003)

Once the stochastic matrix is obtained using the neural network, we determine the entropy rate by using the following equation:

$$H(P) = -\sum_{i=1}^{N} \mu_i \sum_{j=1}^{N} p_{ij} \log_N p_{ij} \tag{2}$$

Where, p_{ij} are the entries of the matrix P and μ represents the stationary distribution. This stationary distribution can be obtained using:

$$\mu_i = \frac{1}{N} \sum_{k=1}^{N} p_{ki}, \quad \text{for each } i = 1, ..., N \tag{3}$$

In order to normalize the entropy, we use the logarithm base N, where N is the size of the *n*-grams vocabulary. Thus, $H(P)$ can take values from 0 to 1. A value close to 1 would represent higher uncertainty in the sequence of *n*-grams within the words in a certain language, i.e, less predictability in

the word formation processes. The calculation of the entropy rate can be summarized in the following steps:

1. For a given corpus, divide every word into its character *n*-grams. A vocabulary of size *N* (the number of *n*-grams) is obtained.
2. Calculate the probability of transitions between *n*-grams, $p_{ij} = p(w_j|w_i)$. This is done using the feedforward neural network described before.
3. A stochastic matrix $P = p_{ij}$ is obtained.
4. Calculate the entropy of the stochastic matrix $H(P) = -\sum_{i=1}^{N} \mu_i \sum_{j=1}^{N} p_{ij} \log_N p_{ij}$.

4. What is the theoretical motivation for this measure

The type-token relationship (TTR) has proven to be a simple, yet effective, way to quantify the morphological complexity of a language using relatively small corpora (Kettunen, 2014). It has also shown a high correlation with other types of complexity measures like paradigm-based approaches that are based on typological information databases (Bentz *et al.*, 2016).

Since TTR is affected by the type and length of the texts, one natural way to make it comparable across languages is to use parallel corpora. TTR has been used for comparing the morphological productivity and the degree of syntheticity and analyticity between languages using parallel corpora (Kelih, 2010; Mayer *et al.*, 2014).

One important underlying intuition of this type of approach is that complexity depends on the morphological system of a language, like its inflectional and derivational processes. A very productive system will produce a lot of different word forms. From a linguistic perspective, Bybee (2010: 9) affirms that "the token frequency of certain items in constructions [i.e., words] as well as the range of types [...] determines representation of the construction as well as its productivity".

Ackermand and Malouf (2013) highlight two different dimensions of morphological complexity: the ennumerative (e-complexity) that focuses on delimiting the inventories of language elements (number of morphosyntactic distinctions and how they are encoded in different word forms); and the integrative complexity (i-complexity) that "measures the (inter)predictability among word forms — i.e., it reflects the ways that the enumerative ingredients cataloged by E-complexity are organized" (Malouf and Ackerman, 2019: 1). In this sense, TTR would fit into the e-complexity dimension, while the entropy rate of a subword language model is closer to the i-complexity.

Entropy is a measure of unpredictability: an event with low probability carries more information. In linguistics, entropy has been used to measure the complexity of morphological systems (Blevins, 2013; Ackerman and Malouf, 2013; Baerman, 2012). Our method aims to reflect how predictable/regular are morphological processes by measuring the entropy of a neural language model trained over a corpus. A morphological process can be unpredictable due to several factors, e.g., a) the process is unproductive (for example, a derivation is less productive than inflection); b) there exists allomorphy; c) a complex system of inflectional classes; d) cumulative (like *portmanteau*) or empty (zero morphs) patterns; e) there exists suppletive patterns.

Statistical language models are a common tool in NLP for estimating a probability distribution over sequences of words. We chose to model this distribution over sequences of *n*-grams of characters within a word, specifically trigrams, since there is evidence that these units encode morphological properties (Baayen, Chuan and Blevins, 2018; Vania, 2017). However, we tried also unigrams.

We conjecture that some languages may produce a lot of different word forms, for example a highly polysynthetic language may seem complex in one dimension (TTR), however, the same language can be quite regular (low entropy). We can think in the opposite case, a language with poor inflectional morphology may have low TTR, however, it may have many suppletive/irregular phenomena, this will cause the entropy go higher. Due to this, we decided not only to estimate the morphological complexity using two different approaches, but to combine (average) the resulting rankings (see Section 6).

5. What the advantages and the drawbacks of this measure are

Advantages:
Our approach is corpus based, this represents a relatively easy and reproducible way to quantify complexity without the strict need of linguistic annotated data.

We try to capture two dimensions of morphological complexity, i.e. take into account not only the productivity of morphological processes in a language but also the predictability of those morphological processes.

Drawbacks:
We believe that our method of entropy of a subword language may be specially suitable for concatenative morphology. There may be several processes that add morphological complexity to a language that are not being taken into account. For example, adding a tone in tonal languages. Some morphological phenomena, like stem reduplication, may seem quite easy from a speaker perspective; however, if this stem is not frequent in the corpus, the language model may find it difficult to predict.

6. If you propose several measures, say explicitly how you are labelling each one in the csv template.

The following labels were used in the template:

1. **H1gram:** Entropy rate using unigrams: (goes from 0 to 1; where 1 is the most complex)
2. **H3gram:** Entropy rate using trigrams: (goes from 0 to 1; where 1 is the most complex)
3. **TTR:** Type-token relationship (goes from 0 to 1; where 1 is the most complex)

Combined rankings:
Once the above measures are obtained, we ranked the languages according to each type of complexity, i.e., rankings that go from 1 (most complex) to 49 (less complex). We average these rankings in order to combine the different complexity dimensions. Finally, we apply the inverse function to the average in order to be consistent with the complexity scale (0 for the least complex. 1 for the most complex):

4. **TTR+H1:** TTR ranking averaged with H_1gram ranking (goes from 0 to 1; where 1 is the most complex)
5. **TTR+H3** : TTR ranking averaged with H_3gram ranking (goes from 0 to 1; where 1 is the most complex)
6. **TTR+H1+H3:** TTR ranking averaged with H_1gram ranking and H_3gram ranking (goes from 0 to 1; where 1 is the most complex)

7. References

Ackerman, F. and Malouf, R. (2013). "Morphological organization: The low conditional entropy conjecture". *Language*, 89(3): 429–464.

Baayen, R. H., Chuang, Y. Y., and Blevins, J. P. (2018). "Inflectional morphology with linear mappings". *The Mental Lexicon*, 13(2): 230-268.

Baerman, M., Brown, D., & Corbett, G. G. (2010). *Morphological complexity: a typological perspective*. Surrey: University of Surrey, MS. Online: http://www. morphology. surrey. ac. uk/Papers/Morphological_complexity. pdf.

Baerman, M. (2012). "Paradigmatic chaos in nuer". *Language*, 88(3):467–494

Bentz, C., Soldatova, T. Koplenig, A. and Samardzić, T.(2016). "A comparison between morphological complexity measures: typological data vs. language corpora". *Proceedings of the workshop on computational linguistics for linguistic complexity*: 142-153.

Blevins, J. P. (2013). "The information-theoretic turn". *Psihologija*, 46(4):355–375

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Kelih, E. (2010). "The type-token relationship in slavic parallel texts". *Glottometrics*, 20(1):1–11.

Kettunen, K. (2014). "Can type-token ratio be used to show morphological complexity of languages?". *Journal of Quantitative Linguistics*, 21(3): 223–245.

Mayer, T., Wälchli, B., Rohrdantz, C., and Hund, M. (2014). "From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets". *Language Processing and Grammars. The role of functionally oriented computational models*: 13–38

Miestamo, M. (2008). "Grammatical complexity in a cross-linguistic perspective". *Language complexity: Typology, contact, change*: 23–41.

Montermini, F. and Bonami, O. (2013). "Stem spaces and predictability in verbal inflection". *Lingue e linguaggio*, 12(2):171–190.

Simon, H. A. (1996). *The architecture of complexity*. Cambridge: MIT Press.

Sinnemäki, K. (2011). *Language universals and linguistic complexity: Three case studies in core argument marking*. PhD dissertation. Helsinski: University of Helsinki.

Vania, C. and Lopez, A. (2017). "From characters to words to in between: Do we capture morphology?". *ACL 2017 Anthology*.